Is the Number Needed-to-Treat (NNT) a key marker of therapeutic efficiency? The GLP-1 receptor agonists as example

L. Monnier[1], C. Colette[1], F. Bonnet[2], D. Owens[3]

1 Institute of Clinical Research, University of Montpellier (France)

2 University Hospital of Rennes, and University of Rennes (France)

3 Diabetes Research Group, Cymru, Swansea University, Wales, UK

Running title: Number Needed-to-Treat with GLP-1 receptor agonists

Word count:

Address for correspondence: Professor Louis Monnier

Institute of Clinical Research, 341 avenue du doyen Giraud, 34093 Montpellier cedex 5, France

E-mail: louis.monnier@inserm.fr

Key words: Number Needed-to-Treat, GLP-1 RAs, treatment efficiency, type 2 diabetes

The efficiency of a medication or a therapeutic class is usually assessed from Randomized Controlled Trials (RCTs) by comparing several "hard" primary or secondary outcomes in persons suffering from a specific disease and who are selected on the basis of predefined criteria and further allocated either to active therapy or a placebo. These interventional studies are commonly referenced to as "mega-trials" [1] because generally conducted in large populations of patients as high as 100 000 people. After random assignment, the eligible participants are equally distributed into 2 groups and submitted to a follow-up period of several months or years. The deleterious or beneficial effects of the tested therapy are assessed at end-point by recording and comparing the occurrence of all-cause deaths or "hard" cardiovascular outcomes. To enhance the strength of the statistical analysis, the clinical outcomes are frequently aggregated into composite endpoints such as "Major Adverse Cardiovascular Events" (MACE) that, however, should be examined with a critical view because such strategies for handling the data can artefactually lead to significant results that do not reflect the truth worthy information provided by the statistical analysis of each event accounted separately. The present commentary discusses the results obtained in recent RCTs [2-6] that were conducted with glucagon-like peptide-1 receptor agonists (GLP-1RAs) in type 2 diabetes. Prior to analyzing the data, the main statistical procedures used in interventional trials should be briefly remembered without burdening the reader with the details of mathematical manipulations.

Analyzing the results with a comprehensive approach

The analysis of RCTs is usually conducted by determining the incidence rates also referenced to as the absolute risks (R1 and R2) of a given clinical outcome in 2 randomly groups of participants selected according to whether they were exposed (group 1) or not (group 2) to an active medication. The ratio R1/R2 corresponds to the relative risk (RR) of the clinical event in group 1 compared to group 2. To address the significance of the reduction in the relative risk it is necessary to complete the calculation of the RR with the assessment of its 95% confidence interval (95% CI) aimed at providing an estimate of the dispersion of the RR. The value of the 95% CI could be ideally obtained from 100 RCTs conducted using strictly similar designs. This complex procedure is routinely bypassed by using formulas that allow its assessment from one single trial and with the same accuracy as that provided by 100 trials. The result is considered to be statistically significant when the horizontal bar that illustrates the 95% CI does not cross the vertical line that corresponds to the neutral result defined by a RR equal to 1. This concept can be extended to the Hazard Ratio (HR). For instance, and as shown in figure 1, it appears that the results were found to be significant ($p < 0.05$) for a composite of adverse cardiovascular events in some trials where GLP-1 RAs were compared to a placebo: the semaglutide (SUSTAIN 6 [2]), liraglutide (LEADER [3]), dulaglutide (REWIND [4]) and albiglutide (Harmony Outcomes [6]). In contrast, the results were not statistically significant in EXSCEL [5] designed to compare the extended-release preparation of exenatide with a placebo. The "p" value is usually considered to be significant when $p < 0.05$. However, in many situations a threshold set at 0.005 or even lower [7,8] would be more appropriate in order to establish the statistical significance on stronger basis, especially when large populations are investigated in different settings and countries.

Number Needed-to-Treat for avoiding one adverse event when patients are treated with a given medication

This number, defined by the acronym NNT [9], can be easily determined from the following formula: $100/(R_2-R_1)$. At this point there arises the question as to whether the determinations of relative risks and NNT are complementary or not. Consider two trials (A and B) that have yielded similar results in terms of RR (0.75). In one of the 2 trials (A) the clinical research was applied to 20 000 subjects who were equally and randomly allocated into 2 arms of 10 000 participants (the exposed and non-exposed patients). The investigators found that an adverse event occurred in 600 and 800 participants with the active therapy and placebo, respectively, over the 5-year study period. The calculated absolute risks are 6% and 8% with a RR value of 0.75 (6%/8%). In trial B 1000 subjects were included in each group. The numbers of events were of 240 and 320 with the active therapy and placebo, respectively. The calculated absolute risks were 24% and 32% with a RR value of 0.75 (24%/32%). However the seemingly similar relative risk in the trials A and B can lead to misleading conclusions if the analysis remains restricted to the crude data provided by the RR and its reduction (- 25%). In contrast the interpretation becomes quite different when the analysis is extended to the NNT. In the trial A, the NNT is of 100/8%-6%), i.e. of 50 subjects whereas in the trial B its value is of 100/(32%-24%), i.e. 12 subjects. This result strongly suggests that the treatment used in the trial B is much more efficient than that administered in the trial A, but the study is inconclusive if we limit the analysis to the reduction in the relative risk.

However, the evidence for a difference between the 2 trials can be demonstrated from the calcuiation of the “p” value, a measure of the strength of statistical evidence [10]. A popular approach of the “p” value is to consider that it is statistically significant when < 0.05 and highly significant when < 0.01. Two main factors play a major role in the statistical significance of the RR or HR (i) firstly the horizontal distance between its value and 1 (the vertical line of neutrality) and (ii) secondly the 95% CI: the smaller the confidence interval, the greater the statistical significance. As the 95% CI is inversely related to the number of participants, it can be easily understood that the investigators have a tendency to study large populations and to analyze different composites of adverse outcomes by bringing together as many categories of clinical events as possible. Such procedures help to reinforce the strength of “p” values on the boundaries of statistical significance when populations remain relatively small or when adverse events are analyzed separately. One example is given in figure 1. The Hazard Ratio (HR) for the expanded composite outcomes is lower in SUSTAIN-6 than in Harmony Outcomes (0.74 vs 0.78) but the strength of the “p” value is higher in Harmony Outcomes than in SUSTAIN-6 (p = 0.0005 vs p = 0.002). This apparent discrepancy is explained by the fact that the 95% CI is smaller in Harmony Outcomes than in SUSTAIN-6: (0.69-0.90) vs (0.62-0.89), the difference being due to the larger number of participants enrolled in Harmony Outcomes (n = 9473) compared with SUSTAIN-6 (n = 3297). This explains why despite their limitations [11,12]“mega-trials” with composites of adverse events as primary end-points are presently the “gold standard” for comparing active therapies to a placebo.

The example of GLP-1 Receptor Agonists with duration of action equal to or longer than 24 hours.

As mentioned above 5 good quality RCTs [2-6] conducted in type 2 diabetes were selected for testing the impact on cardiovascular outcomes of different GLP-1 RAs and for addressing the following issues: (i) Are these glucose-lowering agents equivalent or not in terms of cardiovascular effects? (ii) Is the calculation of the NNT useful or even mandatory to assert on strong basis the potential differences between the analogs? Does it exist a threshold for the NNT above which the treatment looses its clinical pertinence? The main characteristics of the 5 selected trials are reported in table 1.

All analogs are once-weekly preparations, except the liraglutide, a 24-h acting preparation. All participants were persons with type 2 diabetes. The treatment with the GLP-1 receptor agonists was administered either for primary or secondary preventions of the risk for the development or progression of vascular diabetes complications. Patients enrolled within the context of primary prevention were free of any established symptoms of cardiovascular diseases but all of them were at high cardiovascular risk due to the presence of risk factors such as increased blood pressure and dyslipidaemia.

Effects of GLP-1 receptor agonists on the relative risk for developing adverse events

The results concerning the expanded composite of cardiovascular events considered as a whole and the deaths from cardiovascular cause are shown in figure 1. Most trials indicate a significant reduction in the risk of adverse events. SUSTAIN-6 (semaglutide) [2], Harmony Outcomes (albiglutide) [6] and LEADER (liraglutide) [3] were associated with the highest significant relative risk reduction in terms of expanded composite outcomes. Surprisingly, the results for deaths of cardiovascular causes are only significant with the liraglutide (LEADER) [3]. The broadest discrepancies were observed in SUSTAIN-6 (semaglutide) [2] and Harmony Outcomes (albiglutide)[6] because these two studies showed a highly significant reduction for the expanded composite cardiovascular outcomes, whereas these two GLP-1 receptor agonists did not produce any significant changes in the incidence rate of cardiovascular mortality. One of the explanations for this discrepancy can be probably found in the fact that the composite outcomes are not similarly defined in all RCTs. In SUSTAIN-6 [2], Harmony Outcomes [6] and LEADER [3], the investigators included 6 categories of adverse cardiovascular events into the composite outcomes: the deaths from cardiovascular causes + non-fatal myocardial infarctions + non-fatal strokes + revascularizations + hospitalizations for unstable angina pectoris + heart failure. These studies were the most positive (figure 1). In contrast, negative or poorly positive results were observed in EXSCEL [5] and REWIND [4] (figure 1), respectively, when the revascularizations and intercurrent hospitalizations for cardiovascular purposes were excluded from composite outcomes. Such observations point out the difficulties to select appropriate composites because their statistical significance can be rendered very unstable according to whether the choice has been or not properly done.

Effects of GLP-1 receptor agonists on the Number Needed- to- Treat

Does the estimation of the NNT permit to gain better insights into the questions that have been raised by the analysis of relative risks? If we limit our concern to the composite outcomes (table 2) the number of subjects to be treated for avoiding a cardiovascular event in one of them steadily decreases and thus becomes more and more pertinent across a stepwise decrement from EXSCEL (NNT = 125) to SUSTAIN 6 (NNT = 17) with intermediary steps corresponding to REWIND (n = 71), LEADER (NNT = 42) and Harmony Outcomes (NNT = 27). Using the same procedure the tabulation order for deaths from cardiovascular cause is somewhat different, the best NNT being observed for LEADER (NNT = 77) (table 2). The other studies do not provide convincing results because the NNTs were above 100 subjects. Therefore the NNT seems to indicate that GLP-1 RAs are not equally efficient albeit there is an ongoing debate on an overall class effect [13,14].

The argument for using NNT as a marker of therapeutic efficiency despite limitations due to the duration of follow-up

The question that arises from the NNT is to know as to whether there exists a relationship between the NNT and the level of the "p" value calculated from the RR or HR. Using the data provided by the 5 RCTs, the relationship between these two parameters can be depicted by an hyperbolic function (figure 2) the NNT becoming greater and greater when the strength of the "p "value becomes less and less significant. The coefficient of determination (R^2 = 0.68) of the relationship is highly significant: $p < 0.0001$. In addition, the "p" values 0.05 and 0.005 correspond to NNT of 40 and 80 subjects, respectively. If we consider that a treatment becomes justified when the "p" value is < 0.005, consequently we suggest that the relevance for a treatment with GLP-1 RAs should be based o a NNT < 40. Therefore it appears that only 3 preparations fulfill this condition: the semaglutide (NNT = 17), albiglutide (NNT = 27) and liraglutide (NNT = 42). A NNT between 40 and 80 and *a fortiori* > 80 raises the question of the pertinence of treatments with GLP-1 RAs. However it should be noted that the NNT has to be aligned on the duration of treatment in order to be appropriately interpreted.

At the end of this commentary article, the following main conclusions can be drawn:

- The investigators of RCTs should provide all the data related to the analysis of the results in order that everyone might be able to do additional calculations such as those of NNT
- The determinations of relative risks (Risk or Hazard Ratios) and the significances of their changes expressed as levels of the "p" value are mandatory but remain insufficient
- The NNT permits to capture additional information that are likely more meaningful than those provided by the RR, HR and 95% CI, which sometimes are pet peeves for some healthcare professionals who are unfamiliar with the use of statistics
- The definition of the composite end points of adverse events in the different trials should be more carefully harmonized in order that the data of RCTs become easily comparable [13,15].

Funding and disclosure of interest:

No financial support was used in the preparation and writing of this article.

The authors declare that they have no competing interest.

References

[1] Charlton BG. Mega-trials: methodological issues and clinical implications. J R Coll Physicians Lond 1995;29:96-100

[2] Marso SP, Bain SC, Consolli A et al for the SUSTAIN-6 Investigators Semaglutide and cardiovascular outcomes in patients with type 2 diabetes. N Engl J Med 2016;375:1834-1844

[3] Marso SP, Daniels GH, Brown-Frandsen K et al, on behalf of the LEADER Trial Investigators. Liraglutide and cardiovascular outcomes in type 2 diabetes. N Engl J Med 2016;375:311-322

[4] Gerstein HC, Colhoun HM, Dagenais GR et al. Dulaglutide and cardiovascular outcomes in type 2 diabetes (REWIND): a double-blind, randomised placebo-controlled trial. Lancet 2019;394:121-130

[5] Holman RR, Bethel MA, Mentz RJ et al, for the EXSCEL Study Group. Effects of once-weekly exenatide on cardiovascular outcomes in type 2 diabetes. N Engl J Med 2017;377:1228-1239

[6] Hernandez AF, Green JB, Janmohamed S et al, for the Harmony Outcomes committees and investigators. Albiglutide and cardiovascular outcomes in patients with type 2 diabetes and cardiovascular disease (Harmony Outcomes): a double-blind, randomised placebo-controlled trial. Lancet 2018. DOI:https://doi.org/10.1016/SO140-6736(18)32261-X

[7] Monnier L, Colette C, Halimi S. Le seuil du « p » statistique à 0,05 est-il fiable ou non ? Telle est la question. Médecine des maladies Métaboliques 2018 ;12 :671-678

[8] Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond ”p<0.05”. Am Stat 2019;73(Suppl1):1-19

[9] Laupacis A, Sackett DL, Roberts RS. An Assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988;318:1728-1733

[10] Ware JH, Mosteller F, Delgado F et al. P values. In: Medical Uses of Statistics (2nd edition). John C Bailar III and Frederick Mosteller NEHM books, Boston,Massachussetts 1992:181-200

[11] Del Prato S. Megatrials in type 2 diabetes. From excitement to frustration? Diabetologia 2009;52:1219-1226

[12] MonnierL, Colette C, Schlienger J-L, Bauduceau B, Owens DR. Glucocentric risk factors for macrovascular complications in diabetes: Glucose “legacy”and “variability”- what we see, know and try to comprehend. Diabetes Metab 2019;45:401-408

[13] Bethel MA, Patel RA, Merrill P et al. Cardiovascular outcomes with glucagon-like peptide-1 receptor agonists in patients with type 2 diabetes: a meta-analysis. Lancet Diabetes Endocrinol 2018;6:105-113

[14] Scheen AJ. GLP-1 receptor agonists and cardiovascular protection: A class effect or not ? Diabetes Metab 2018;44:193-196

[15] Monnier L, Colette C, Schlienger J-L, Halimi S. Les méta-analyses en recherche clinique : forces et faiblesses. Médecine des maladies Métaboliques 2019 ;14 (à paraitre, sous presse)

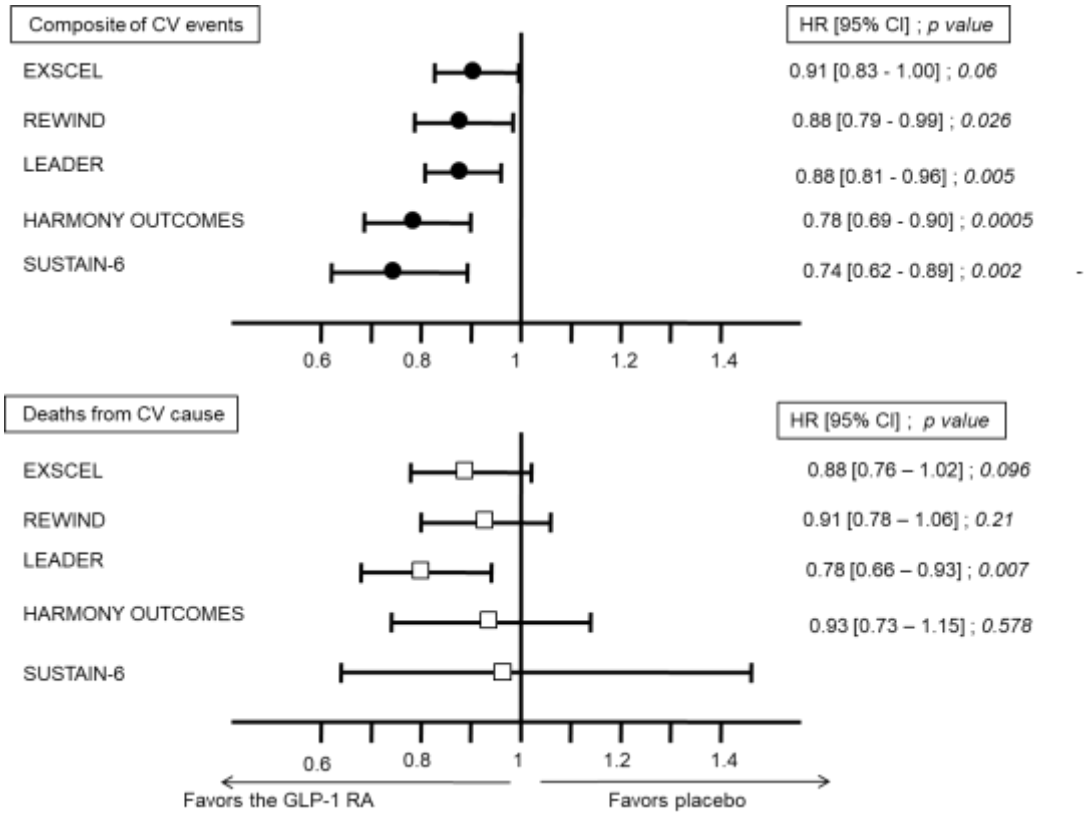

Figure 1: Hazard Ratio (HR) and "p" values in 5 Randomized Controlled Trials (RCTs) aimed at examining cardiovascular outcomes when different GLP-1 receptor agonists were compared with a placebo in type 2 diabetes. The HRs were indicated for an expanded composite of cardiovascular events (upper part of the figure) and for deaths of cardiovascular cause (lower part of the figure)

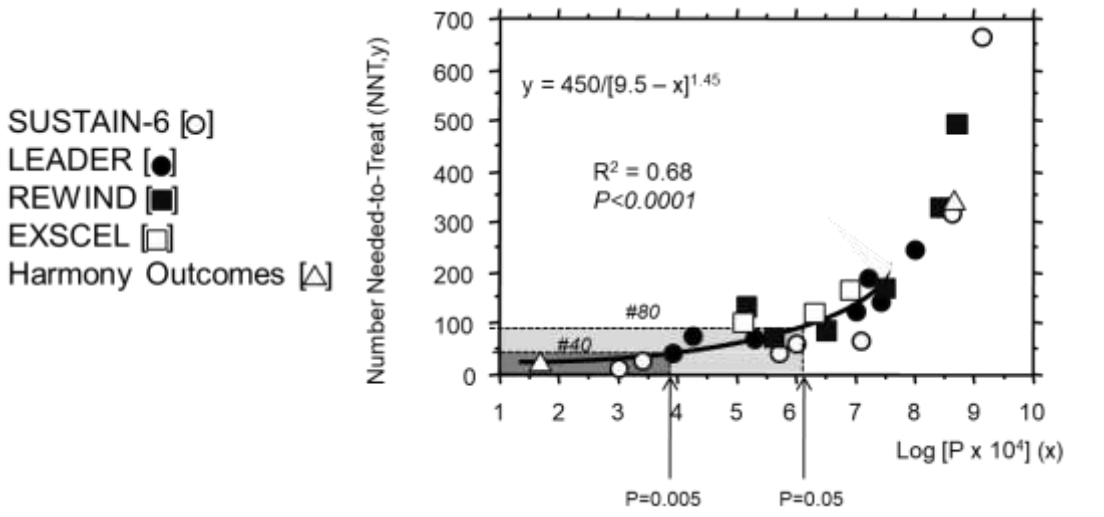

Figure 2: Relationship between "p" values and Hazard Ratios (HR) for different cardiovascular events (X axis) and the Number Needed-to-Treat (NNT, Y axis) for avoiding one cardiovascular event in one of the patients treated with a GLP-1 receptor agonist. To facilitate the illustration the "p" values were converted into Log ($px10^4$). The "p" threshold set at 0.05 and 0.005 corresponds to NNT of 40 and 80, respectively. The grey zones (dark or light) correspond to the different objectives: NNT < 40 or NNT between 40 and 80, respectively.

RCTs	Age (years)	GLP-1 RA tested (doses)	N° of patients	Prevention	Mean duration (years)	Reduction in HbA1c (%)
SUSTAIN-6	64.6	Semaglutide (0.5 or 1.0 mg per week)	3297	Primary or secondary	2.0	- 0.66 % (0.5 mg/week) - 1.05 % (1.0 mg/week)
Harmony Outcomes	64.1	Albiglutide (30-50 mg per week)	9463	Secondary	1.6	- 0.52 %
LEADER	64.3	Liraglutide (maximal tolerated dose or 1.8 mg per day)	9340	Primary or secondary	3.8	- 0.40 %
REWIND	66.2	Dulaglutide (0.75 or 1.5 mg per week)	9901	Primary or secondary	5.4	- 0.61 %
EXSCEL	62.0	Extended-release Exenatide (2 mg per week)	14752	Primary or secondary	3.2	- 0.53 %

Table 1: Main characteristics of the 5 RCTs involving a comparison between patients with type 2 diabetes according to whether they were exposed or not to a treatment with different GLP-1 receptor agonists. The secondary prevention was defined by the presence of at least one cardiovascular coexisting condition. The primary prevention was characterized by the lack of any evidence of cardiovascular disease at baseline. However due to the presence of risk factors, these subjects were at high risk of cardiovascular outcomes.

RCTs	HR ; *P*	NNT	RCTs	
SUSTAIN-6	0.74 ; *0.002*	17	SUSTAIN-6	
Harmony Outcomes	0.78 ; *0.0005*	27	Harmony Outcomes	Significant
LEADER	0.88 ; *0.005*	42	LEADER	
REWIND	0.88 ; *0.026*	71	REWIND	
EXSCEL	0.91 ; *0.06*	125	EXCEL	NNT>80

RCTs	HR ; *P*	NNT	RCTs	
LEADER	0.78 ; *0.007*	77	LEADER	Significant
EXSCEL	0.88 ; *0.096*	167	EXCEL	
REWIND	0.91 ; *0.21*	167	REWIND	NNT>80
Harmony Outcomes	0.93 ; *0.578*	333	Harmony Outcomes	
SUSTAIN-6	0.98 ; *0.92*	666	SUSTAIN-6	

Table 2: Stratification of the cardiovascular risk by taking into account the HR and NNT in 5 RCTs aimed at comparing patients with type 2 diabetes according to whether they were treated or not with different GLP-1 receptor agonists. Upper part: statistical risk for a composite of cardiovascular outcomes. Lower part: statistical risk for cardiovascular mortality. CRTs = Controlled Randomized Trials. NNT = Number Needed-to-Treat for avoiding one adverse event in one of the patients treated with the GLP-1 receptor agonist.